

UNIVERSITI SAINS MALAYSIA



UNIVERSITI SAINS MALAYSIA

**DISTRIBUTION OF ALLELE FREQUENCIES OF SHORT
TANDEM REPEATS (STRs) D16S539, D7S820 AND D13S317 IN
RANDOM TEMIAR POPULATION GROUP OF MALAYSIA**

**DISSERTATION SUBMITTED IN PARTIAL FULFILMENT
FOR THE DEGREE OF BACHELOR OF SCIENCE (HEALTH) IN
FORENSIC SCIENCE**

KHOR KOK PIN

**SCHOOL OF HEALTH SCIENCES
UNIVERSITY SAINS MALAYSIA
16150 KUBANG KERIAN KELANTAN
MALAYSIA**

2009

CERTIFICATE


This is to certify that the dissertation entitled
**Distribution of Allele Frequencies of Short Tandem Repeats (STRs), namely D16S539,
D7S820 and D13S317 from Random Temiar Population Group of Malaysia**

is the bonafide record of the research work done by

Mr. Khor Kok Pin

during the period from January 2009 to April 2009 under my/ our supervision

Signature of Supervisor:


.....

(Mr. S. Panneerchelvam)

School of Health Sciences, University of Science Malaysia, 16150 Kubang Kerian
Kelantan, Malaysia

Signature of Co-supervisor:

.....

(Prof. Norazmi Mohd. Nor)

School of Health Sciences, University of Science Malaysia, 16150 Kubang Kerian
Kelantan, Malaysia

ACKNOWLEDGEMENT

I hereby wish to grab the opportunity to thank for the several parties who have been kindly lend out their helping hand in supporting this research study. Without the aids from all of you, this study could not have been completed on time. First and foremost, I would like to send my utmost gratitude toward my research supervisor, Mr. Panneerchelvam who has given so much invaluable suggestions and opinions whenever I faced difficulties and clag in completing this research. Thanks for his patience while guiding me, as well as giving me the opportunity to experience this research study.

Next, I wish to thank Ms. Allia Shahril who had assisted me in running the research protocols. She had taught me the excellent skills of running the PCR and electrophoresis procedures, as well as providing me the brilliant knowledge in counteracting and troubleshooting difficulties faced in this research. Her constructive comments are much appreciated. Not forgotten to thank her and her colleague, Ms. Wahida for spent time in collecting the buccal swab samples from Temiar population.

Third, I would like to express my acknowledgement toward officers of Unit Kemudahan Makmal (UKM) and the INFORMM for willingly preparing the reagents and facilities that utilized in this research. Specially thanks for the INFORMM officer, Mr. Bat in providing the E Pure water which is very important in preparing the polyacrylamide gel for electrophoresis.

Last but not the least, I would like to thank Prof. Ahmad bin Zakaria, the Dean of PPSK, Prof. Sayed Mohsin Sayed Jamallulail, Dean of Research and Development, Prof. Norazmi Md. Noor for providing me the opportunity to experience the research, as well as lecturers of forensic science course and other persons who assisted me in completing this research.

TABLE OF CONTENT

| | |
|---|-----------|
| Acknowledgement | iii |
| Table of contents | v |
| List of tables | vii |
| List of figures | viii |
| Abstract | ix |
| | |
| CHAPTER ONE: INTRODUCTION | 1 |
| CHAPTER TWO: LITERATURE REVIEW | 3 |
| 2.1 Chromosome and Deoxyribonucleic acid (DNA) | 3 |
| 2.2 DNA Profiling | 6 |
| 2.2.1 Restriction Fragment Length Polymorphism (RFLP) | 7 |
| 2.2.2 Polymerase Chain Reaction | 9 |
| 2.2.3 Short Tandem Repeat | 11 |
| 2.3 DNA database | 13 |
| 2.4 Population database | 15 |
| 2.5 Objectives of the study | 17 |
| CHAPTER THREE: MATERIAL AND METHODS | 18 |
| 3.1 Sample source | 18 |
| 3.2 Material | 19 |
| 3.2.1 Equipment and apparatus | 19 |

| | |
|--|-----------|
| 3.2.2 Chemical and reagents | 19 |
| 3.3 Methodology | 20 |
| 3.3.1 Quantification of DNA | 20 |
| 3.3.2 PCR Amplification | 21 |
| 3.3.3 Polyacrylamide gel electrophoresis | 23 |
| 3.3.3.1 Reagent preparation..... | 23 |
| 3.3.3.2 Glass plate preparation methodology | 24 |
| 3.3.3.3 Electrophoresis using SA 32 sequencing electrophoresis apparatus | 24 |
| 3.3.3.4 Silver Staining | 25 |
| 3.3.3.5 Data compilation..... | 27 |
| 3.3.4 Safety and precautions | 27 |
| CHAPTER FOUR: RESULT | 29 |
| 4.1 Determination of STRs by silver staining method | 30 |
| 4.2 Data comparison with other populations | 32 |
| CHAPTER FIVE: DISCUSSION | 46 |
| CHAPTER SIX: CONCLUSION | 49 |
| CHAPTER SEVEN: REFERENCES | 50 |

LIST OF TABLES

| | |
|---|----|
| Table 1: Information on Silver STR ^a III triplex | 12 |
| Table 2: The equipments required in this research with their supplier or company | 19 |
| Table 3: The chemicals required in this research with their supplier or company | 19 |
| Table 4: The volumes of each PCR master mix components | 22 |
| Table 5: PCR amplification protocol for STRIII multiplex | 22 |
| Table 6: The reagents and its components that required in running the polyacrylamide gel electrophoresis | 23 |
| Table 7: Silver staining protocol | 26 |
| Table 8: The detrimental effects of reagents used in this research onto human body | 28 |
| Table 9: Distribution of Genotype Frequency in Temiar population for STR III loci | 31 |
| Table 10: Distribution of Allele Frequency in Temiar population for STR III loci | 31 |
| Table 11: Distribution of allele frequencies for D16S539 in other populations | 32 |
| Table 12: Distribution of allele frequencies for D7S820 in other populations | 36 |
| Table 13: Distribution of allele frequencies for D13S317 in other populations | 41 |

LIST OF FIGURES

| | |
|--|-----------|
| Figure 1: Structure of deoxyribonucleic acid (DNA) elucidated by Watson and Crick in 1953 | 4 |
| Figure 2: Level of packing of mitotic chromosome in human body | 5 |
| Figure 3: Representative figure of PCR amplification | 10 |
| Figure 4: Preparatory agarose gel showing extracted DNA from 28 samples | 21 |
| Figure 5: STRs results of 12 samples with L as the allelic ladder for reference purpose | 30 |

ABSTRACT

Polymerase Chain Reaction based Short Tandem Repeat (PCR-STR) method provides an unprecedented DNA profiling technique for forensic community in identification and individualization purposes. In the analysis of biological material using DNA profiling technique, the information regarding the distribution of allele frequency in a defined population is mandatory. This particularly necessary in paternity testing that involves compilation of Paternity Index (PI), in which the allele frequency is required to indicate whether the alleged father is the true biological father of the child. Validated STR loci have been analysed extensively in various populations for forensic application. In this research study, population genetic studies were carried out for three STR loci, namely D16S539, D7S820 and D13S317 for Temiar population by using a multiplex PCR reaction followed by 4% polyacrylamide gel electrophoresis (PAGE) and silver staining. The heterozygosity (H) for these three loci, i.e. D16S539, D7S820 and D13S317 are 0.9568, 0.9132 and 0.7427 respectively. Power of Discrimination (PD) for these three loci are 0.9135 (D16S539), 0.9331 (D7S820) and 0.9050 (D13S317); Power of Exclusion (PE) was found to be 0.5192 (D16S539), 0.6034 (D7S820) and 0.4053 (D13S317) respectively. The Cumulative Discrimination Power (CDP) for the three STRs for Temiar population is 0.99903.

CHAPTER ONE INTRODUCTION

The advances in techniques of mapping and sequencing the human genome increased the advantages of deoxyribonucleic acid (DNA) typing in the forensic science. The discoveries of various molecular biology approaches such as restriction fragment length polymorphism (RFLP), polymerase chain reaction (PCR) and short tandem repeat (STR) have revolutionized human identification in forensic science. The most widely used DNA profiling, i.e. STR DNA typing has gained popularity since 1993 in forensics due to several significant advantages.

The PCR amplified STR typing technology enables scientists to amplify the number of DNA target molecules from 1ng of DNA template. This is particularly important in forensic science since the crime sample collected always exhibits degraded DNA. Typically as little as 1ng of genomic DNA can be used to produce a full STR profile of an individual, and it provides superior advantage over other methods (single locus probe and multilocus probe) which require at least 10 µg of high molecular weight DNA. Automation of STR technology is less time consuming task than SLP profiling technique, which allows more STR loci to be identified and provide higher discrimination thereby individualization (Sambrook *et al.*, 1989; Erlich, 1989).

The STR loci, also known as microsatellite DNA are characterized by extreme polymorphism levels both within and among the population of individuals (Rowold, 2003).

These microsatellite loci feature high level and stable polymorphisms, uniform chromosomal distribution as well as short sequence length, which are desirable and advantageous in forensic DNA analysis and sequencing (Hammond *et al.*, 1994; Kimpton *et al.*, 1993), intra-species phylogenetic reconstructions (Bowcock *et al.*, 1994; Jorde *et al.*, 1995), paternity testing (Hammond *et al.*, 1994) and forensic analysis (Edward *et al.*, 1991 & 1992).

In this research, the allele frequencies of three STR loci, i.e. D7S820, D13S317 and D16S539 were established for the Temiar population in Malaysia by using 101 saliva samples collected from random Temiar population group. Malaysia is a country where various indigenous groups known as Orang Asli living in different states of the Peninsular Malaysia. These *Orang Asli* which is also known as “aboriginal people” consists of three main tribal groups, which are Semang, Senoi and Proto-Malay. Demographic data in the year of 2000 showed that there are 113,541 Orang Asli individuals from the 18 sub-ethnic groups divided from the three main tribal groups (Centre for Orang Asli Concern). The Orang Asli contributes 0.5% of the total Malaysian population (Alebreto, 2000).

Temiar is one of the sub-ethnic groups classified under the Senoi tribal group with a population of approximately 20,000 individuals (Joshua Project, 1999). They are numerous in Perak, Kelantan, Pahang, Selangor, and Negeri Sembilan states with different or alternative names such as Grik, Pie, and Northern Sakai. At present, Malaysia is the only country in the world where this ethnic group could be found.

CHAPTER TWO LITERATURE REVIEW

2.1 Chromosome and Deoxyribonucleic acid (DNA)

The human body consists of three billion cells, which are the basic units of life. All living cells have a nucleus, which contains DNA as the biological blueprint of heritable characteristics. The role of DNA as the vehicle of generational transference of heritable traits is well defined in 1944 by Oswald Avery and his colleagues (Avery, Macleod and McCarty, 1944). In 1953, James Watson and Francis Crick successfully elucidated the double helix structure of DNA molecule.

DNA consists of two strands of nucleotides that made up of a nitrogenous base, a pentose sugar (5-carbon sugar) and a phosphate group. The description of DNA as elucidated by Watson and Crick in 1953 is as follows:

- (a) Two long polynucleotide chains around a central axis forming a right handed double helix
- (b) The two strands are anti-parallel and bases of both chains are paired to each other by formation of hydrogen bonds (A binds to T, G binds to C)
- (c) Presence of alternating major grooves and minor grooves along the axis at any segments of DNA molecule
- (d) Diameter of DNA measured 2.0nm and each complete turn of helix is 3.4nm

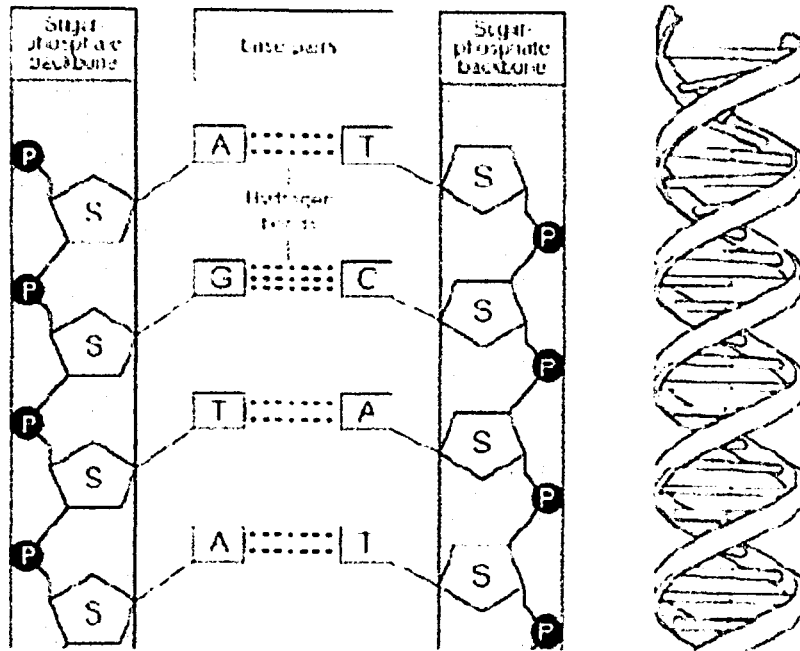


Figure 1: Structure of deoxyribonucleic acid (DNA) elucidated by Watson and Crick in 1953

The human genome project estimated that human genome contains approximately 30,000 genes. These gene sequences encoded by 10% of the genome, i.e. about 300 million bp. The non-coding sequence, comprising of 2.7 billion bp does not contain genetic information related to specific body functions. Approximately 3% of the human genome is made up of tandem repeat sequences (International Human Genome Sequencing Consortium, 2001), which can be classified into two main groups depending on their size of repeat unit and overall length of repeat sequences. These two groups are mini- and microsatellite DNA.

The nuclear DNA and mitochondria DNA in human cell consists of organized dense packets of DNA and protein which is known as the chromosomes. A chromosome is defined as a single piece of DNA that contains genes, regulatory elements and other

nucleotide sequences. Human genome consists of 22 pairs of autosomal chromosomes and two sex determining chromosomes, with males designated XY and females XX (Tijo and Levan, 1956). There are four types of chromosomes, which are metacentric, submetacentric, acrocentric or telocentric based on the location of centromere. Short arm of chromosome is p arm (p stands for “petite”) while the longer arm is known as q arm.

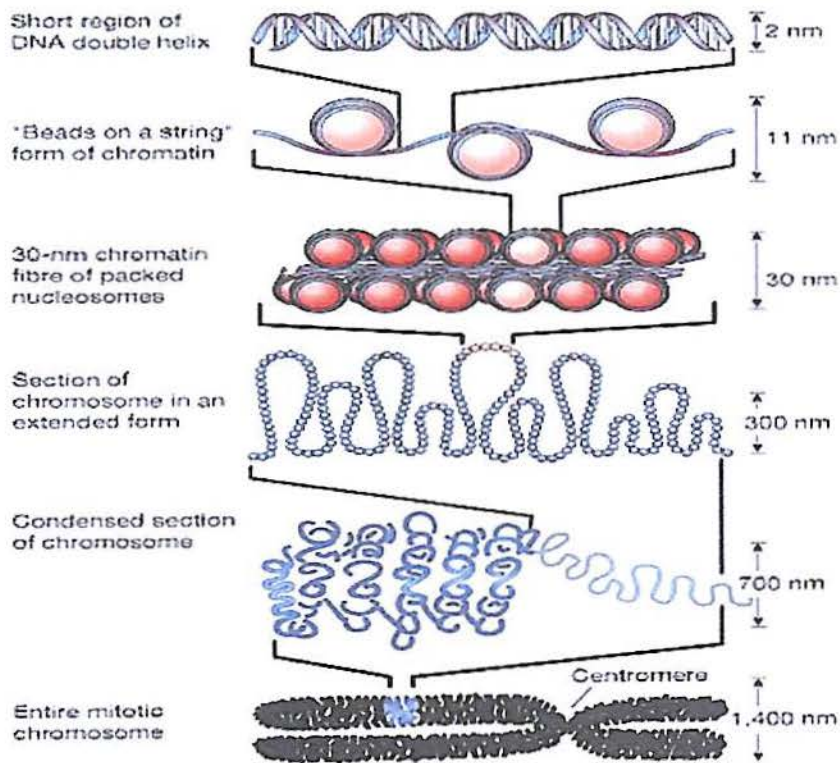


Figure 2: Level of packing of mitotic chromosome in human body

2.2 DNA Profiling

The forensic DNA profiling, also termed as DNA typing or DNA fingerprinting, was originated by Alec Jeffreys in 1985 (Jeffreys and Wilson, 1985). The history of DNA profiling begins in Leicester University of Great Britain when Alec Jeffrey studied the utilization of restriction enzymes and DNA probes in profiling of myoglobin gene cluster. He found the potentiality of using these repetitive regions of genome to individualize biological fluids. To date, the DNA profiling has undeniably become the standard technique in criminal investigations, mostly involved in the analysis of biological stains located at the crime scene, such as semen, blood, saliva, and others.

The prime factor that made DNA profiling a superb tool in forensic investigation prevailing over others was due to its successful analysis of any biological stains that containing nucleated cells (Schneider, 1997). In contrast to the conventional method of blood group and enzyme analysis (e.g. ABO, PGM1, AcP, EsD, Km), the DNA analysis provides higher discrimination power in analyzing the biological stains collected from crime scene. The advantages of DNA typing over other serological methods are being more specific, informative, tissue independent, sensitive and resistant to degradation.

The forensic DNA typing is based on the comparison of polymorphisms of DNA between individuals (Rudin and Inman, 2001). These differences or polymorphisms can be the result from single nucleotide polymorphism (changes in single base pair), or variation of

region length or locus in the genome. Nowadays, short tandem repeat (STR) is the most commercially available automated multiplex systems that used in the forensic caseworks to identify the evidence from the crime scene. However, it's worth to highlight some of the other systems that previously used by the scientists to analyze the sample. The following sections provide brief information in relating to various gene techniques that have been used in forensic laboratory.

2.2.1 Restriction Fragment Length Polymorphism (RFLP)

The DNA analysis first suggested in early 1980's is the RFLP analysis (Jeffreys, 1985). The RFLP analysis can be divided into two main categories, which are multiple locus (MLP) and single locus probe (SLP) based RFLP analysis. These analyses performed by using human minisatellite loci, which have repeat units from 7bp to more than 100bp. The minisatellite loci show very high levels of allele length variability with arrays usually kilobases in length. For example, human GC-rich minisatellites are preferentially found clustered in subtelomeric regions of chromosomes (Royle *et al.*, 1988).

The RFLP analysis had shown great degree of discriminating power if several loci were looked for. This test can easily differentiate two evidences from uncommon sources since there are as many as hundred of variations at each particular locus. However, the RFLP technique requires great amount of good quality DNA from the evidence collected from the crime scene, which is not possible in forensic caseworks since the evidences recovered are often old-aged, degraded and present in limited quantity. Beside that, this technique is

also laborious and non-automated, causing it edged out by other advance automated technology.

In 1984, the MLP system was introduced by Alec Jeffrey, in which the DNA extract is being digested and added two probes. The repeating sequence is then identified through Southern blot analysis to produce a genetic fingerprint of that individual. The DNA fragments with sizes between 4 and 20 kb were used for analysis. This is due to the density of bands below the 4 kb level could not be reliable, while the bands ranged beyond 20kb consisted of a ladder-like series of bands similar to a bar code.

In the MLP system, common 10–15 bp ‘core’ GC-rich sequence that shared between different minisatellite loci allowed it to detect many different minisatellites simultaneously, therefore producing multiband patterns known as ‘DNA fingerprints’. Researchers have established the match probability of these MLPs, using 33.15 MLP probe between unrelated people was estimated at about 3×10^{-11} , while two MLPs (33.15 and 33.6) together gave a value of 5×10^{-19} (Jeffreys, Wilson and Thein 1985). During the early time, MLPs have been used successfully in paternity testing (Jeffreys, Turner and Debenham, 1991) and immigration cases (Jeffreys, Brookfield and Semeonoff, 1985).

After a few years, Single Locus Probing system (SLP) method shone a bright light to the mainstream of forensic science organization. The probes in SLP system identify repeat sequences in individual mini-satellite DNA using high stringency conditions to produce a

maximum two bands per probe (Wong *et al.*, 1987). Variation of repeat numbers provided by single probe was insufficient for individualization of the evidence sample, thus a number of loci were chosen to produce a composite result. In contrast to the MLP system, the advent of SLP system forms the basis of forensic database as the band results obtained can be recorded in numerical form. The size of the bands is estimated through the addition of a control ladder with known molecular size fragments. Single locus probe based analysis is more advantageous when the DNA is degraded or mixed.

2.2.2 Polymerase Chain Reaction

In the early 1990's, the invention of PCR, acronym of Polymerase Chain Reaction (PCR) caused a significant change in the genetic field purposes that enable scientists to amplify template DNA from minimal amounts of extracted DNA (Saiki *et al.*, 1986). The reasons for PCR method being widely used were due to the simplicity of the reaction, fast speed, reliable and extremely sensitive so that small repeating sequences can be detected and amplified. Nowadays, PCR was amenable to automation and allowed to amplify the fragments at several loci in a single reaction.

PCR is an enzymatic processes in which specific region of DNA is replicated over and over again to yield many copies of a particular sequence. The PCR instrument carries out three separate processes to amplify the template DNA into millions of copies, which are denaturation, annealing and extension steps. The one-cycle process is then repeated for about 25-40 times so that huge amount of good quality DNA can be achieved.

The first step performed at 95°C to separate the hydrogen bonds between the DNA strands so that each separated strands can be used as the template for the synthesis of a new DNA strand. After that, the solution is cool down to between 50 and 65°C so that primers can be specifically bind on their complementary sites. The annealed oligonucleotides act as primer for DNA synthesis, since they provide a free 3' hydroxyl group for DNA polymerase. Finally, the solution temperature is raised to 72°C and DNA polymerase will start to incorporate the DNA building blocks onto the free DNA templates.

POLYMERASE CHAIN REACTION

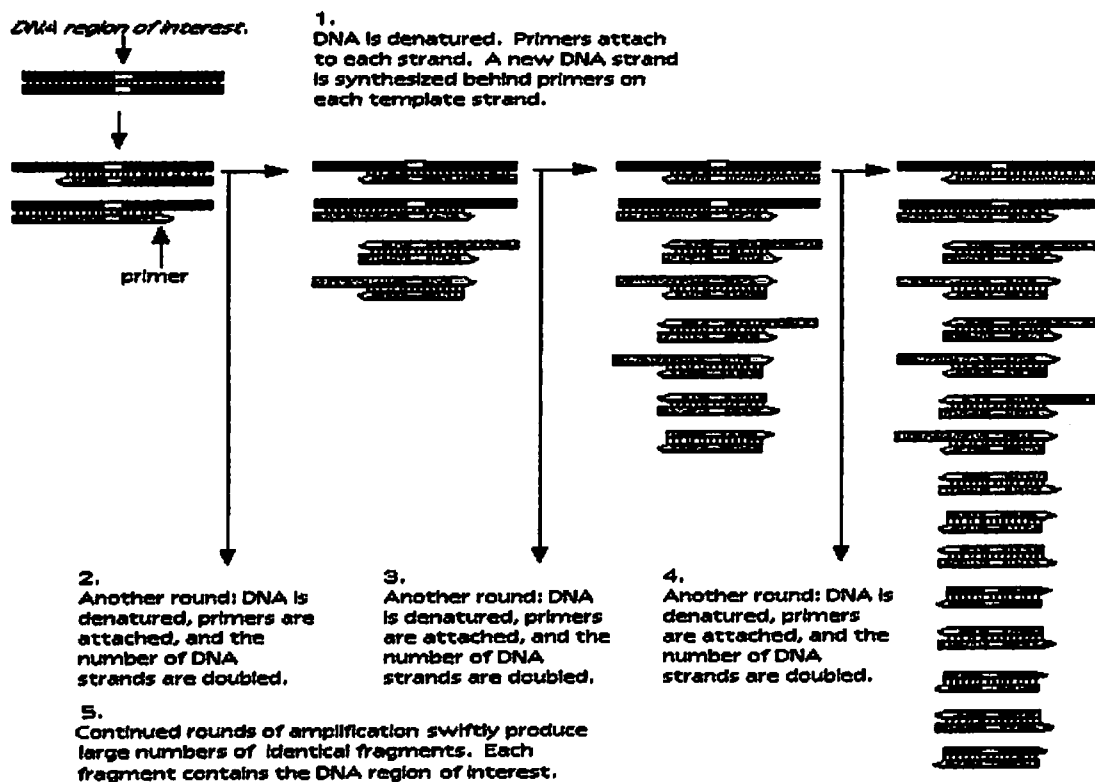


Figure 3: Representative figure of PCR amplification

2.2.3 Short Tandem Repeat

The STR was applied in forensic caseworks to develop the genetic fingerprinting of an individual (Litt and Luty, 1989). It is a class of biology polymorphism in which the repeating sequences are directly adjacent to each other without interspersed elements. STRs are distributed throughout the human genome occurring with a frequency of one locus every 6-10 kb (Beckman and Weber, 1992). Variability of STR is caused by replication slippage during the meiosis process.

Nomenclature of STR is according to length of core repeat units, for examples dinucleotide (two repeat units), trinucleotide (three repeat units), tetranucleotide (four repeat units) and so on. There are three types of STRs, which are simple repeats, compound repeats and complex repeats. Simple repeats are those repeats with similar length and sequence. Compound repeats are made up of two or more simple repeats while complex repeats contain several core repeat blocks of variable unit length (Urquhart et al., 1994).

Genetic typing by using STRs loci become an effective tool in forensic sample identification purpose as well as paternity testing as mentioned earlier. These STRs loci give high statistical discrimination and individualization values (Edward *et al.*, 1991). Generally, short sizes of STRs are preferable in case of degraded samples since the small fragments can be easily amplified compared to larger fragments. STR analyses are faster, easier and more sensitive than RFLP system. Multiplex STR techniques provide better

discriminatory power than RFLP system does, leading the PCR-based STR testing a dominant technique in forensic DNA typing.

In this research study, the distributions of allele frequency of three STR loci, namely D16S539, D7S820 and D13S317 are being studied. These three loci also known as silver STR[®] III triplex. The table below showing the information regarding this multiplexes system.

Table 1: Information on Silver STR[®] III triplex.

| STR locus | Chromosomal location | Range of bases | Number of repeats | Repeat sequence |
|------------------|-----------------------------|-----------------------|--------------------------|------------------------|
| D16S539 | 16q24-qter | 264-304 | 5,8,9,10,11,12,13,14,15 | AGAT |
| D7S820 | 7q11.21-22 | 215-247 | 6,7,8,9,10,11,12,13,14 | AGAT |
| D13S317 | 13q22-q31 | 165-197 | 7,8,9,10,11,12,13,14,15 | AGAT |

2.3 DNA Database

A study in 1994 shown that 60% of those individuals who released from prison for violent offenses were rearrested for a similar offense in less than 3 years time. Subsequently, it was suggested for the co-operation amongst the community or countries in which the results from DNA analysis are exchangeable to facilitate the investigation of forensic cases. The collaboration effort started to establish a program which gathers results from other states or communities, leading the formation of DNA databases.

The emergence of DNA databases provided several important roles, either in assisting the forensic cases or to decrease the number of criminal cases. The database enables the police officer to include or eliminate the suspect easily. The DNA database increased the possibility that serial offenders could be identified after their first or second offenses, thus leading the potential victims in future to be decreased. Beside that, the knowledgeable criminals also realized that they might be nabbed if they repeat criminal offence.

The DNA profiling through analysis of short tandem repeat gained a united agreement in 1992 by members of European DNA Group (EDNAP), which provided a big leap toward forming a systematic and efficient DNA database (Weber & May, 1989). Many forensic scientists from the European countries agreed on the fact that STR would form the basis of the future work. Thus, a series of inter-laboratories exercises were done (Gill *et al.*, 1994).

The first DNA database that involves STR analysis was formed in 1995 at United Kingdom, which results demonstrated the effectiveness in convicting the offended criminals. The total number of personal DNA profiles was amounted to 700,000 by the end of 1999, with around 700 matches achieved each week. Thus, it's not surprising that such an efficient, cost-saving and lesser time-consuming database have gained popularity all over the countries in the world.

In 13 October 1998, FBI officially launched its national DNA database known as Combined DNA Index System (CODIS). The efforts were given out by all 50 states in US to provide DNA sample for the database, leading to a formation of successful database to counteract on crime. The inter-changeability of the DNA profiles amongst the states enhanced the ability of the scientists to locate the offenders. To date, all the 50 states in US had enacted legislation to establish a DNA databank to contain the genetic profiles of individuals convicted of specific crimes.

There are two primary sample indexes in the CODIS system, which are convicted offender samples and forensic casework samples. The CODIS is composed of local, state and national levels. Recently, all the laboratories in the US are converting and storing the DNA profiles in 13 STR loci form, instead of RFLP markers that were used before this.

2.4 Population Database

A population can be defined as a subdivision of a species which hold a community of individuals where mates are usually found. The population of individuals shares a common gene pool and has continuity through time. The measurement of the relative frequency of an allele at a particular locus within a population is termed allele frequency. The population genetics come through the study of these allele frequency distribution and changes under the influence of four evolutionary forces, which are natural selection, genetic drift, mutation and gene flow.

Allele frequency distribution is widely applied when dealing with the paternity testing, which is used to determine whether the tested man is or is not the biological father of the alleged child. Inclusionary result will be reported in a statistical form known as Combined Paternity Index (CPI), which a value of 100 or greater is standardized accepted level to establish the parental rights in most states. Paternity testing is based on the principle that the alleged father (AF) will share the paternal marker with the child for each allele tested. Therefore, exclusionary result was produced if three or more alleles were tested mismatch between the AF and the child.

The inclusionary results were calculated based on the Paternity Index (PI) that compares the likelihood that a genetic marker was passed by the AF to the child over the probability that a randomly selected unrelated man of the similar ethnic background could pass the